IST 652 – Scripting for Data Analysis
Final Project Report

Team members :
Sravan Kumar Mangalagiri (smangala@syr.edu)
Sisira Pathakamuri (spathaka@syr.edu)

Topic of investigation: New York Taxi2022

We are getting Green Taxi datasets from January to the month of august of the year 2022.

NEWYORK TAXI 2022: This table shows the NYC taxi places where it went called zones and which has pick-up and pick -off time and date zones, which includes green taxi trip records published to open by NYC taxi & Limousine Commission.

Green Taxis Data Set:

Field Name	Description
VendorID	A code indicating the LPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
lpep_pickup_datetime	The date and time when the meter was engaged.
lpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride

Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,”because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip
Payment_type	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed on hailed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.
Trip_type	A code indicating whether the trip was a street-hail or a dispatch that is automatically assigned based on the metered rate in use but can be altered by the driver. 1= Street-hail 2= Dispatch

Potential Development tasks:

1. Assess current trends in cash and cashless (Credit Card/Contactless) payments.

2. What characteristics influence passengers' decisions to share rides or go alone on weekdays and weekends? ,Daytime: between 9 a.m. and 5 p.m., or overnight ?

3. Annual patterns for group rides: to determine whether or not the travellers appreciate traveling in groups more. This analysis can help people comprehend the benefits of carpooling and raise their awareness of the pollution caused by vehicle emissions.

4. The volume of reservations for rides at various times of the week in various parts of New York City. And try to find out what influences a higher density of rides in a particular location at a certain time.

5. Provide images of taxi rides in various seasons like in Holidays.

6.What influences how much passengers tip the driver? based on the ride's route, drop-off location, tourism destination, and other factors.

7. Consider the time of day, the location, and how it influences the rate of ride booking when evaluating the ups and downs of surge prices.

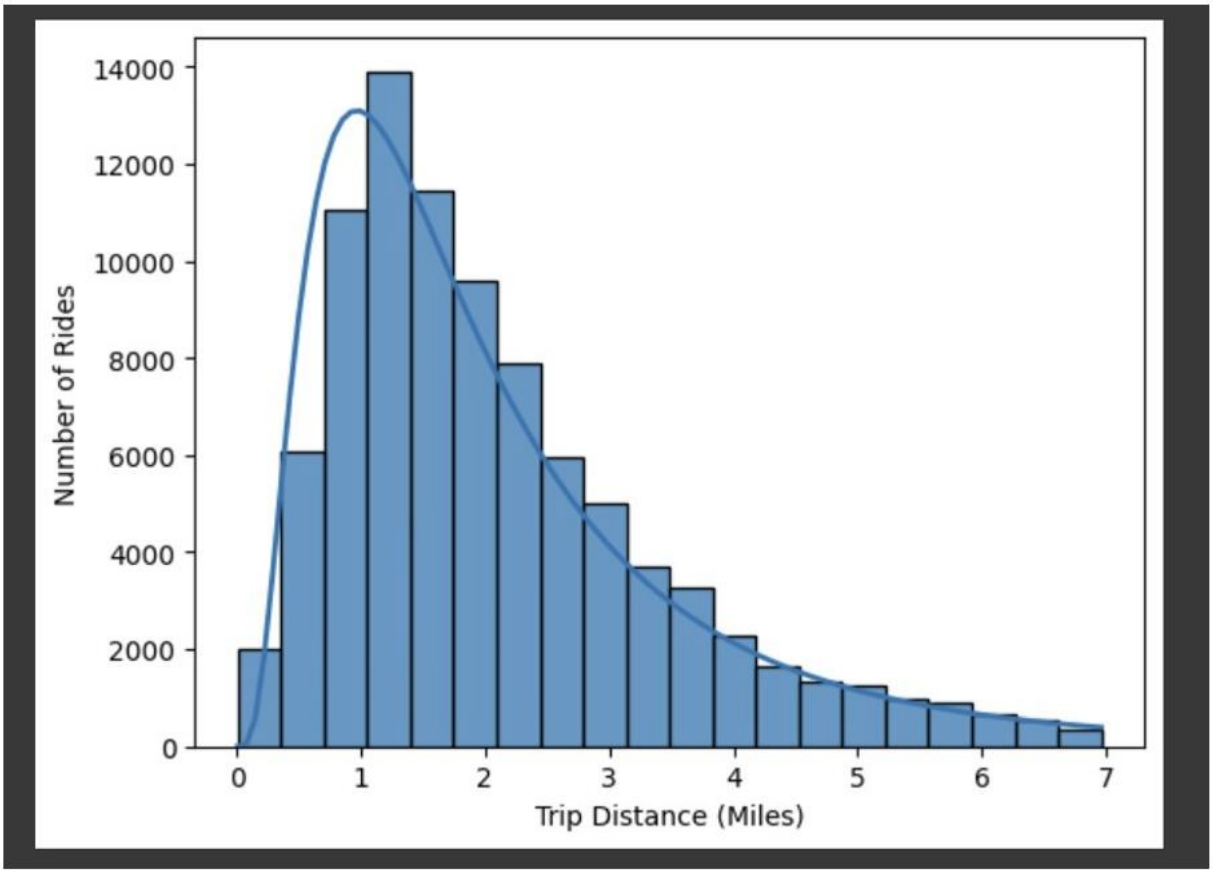
8. Develop visualizations and perform a comparative analysis for each of the aforementioned study topics green cab.

Data Preparation:

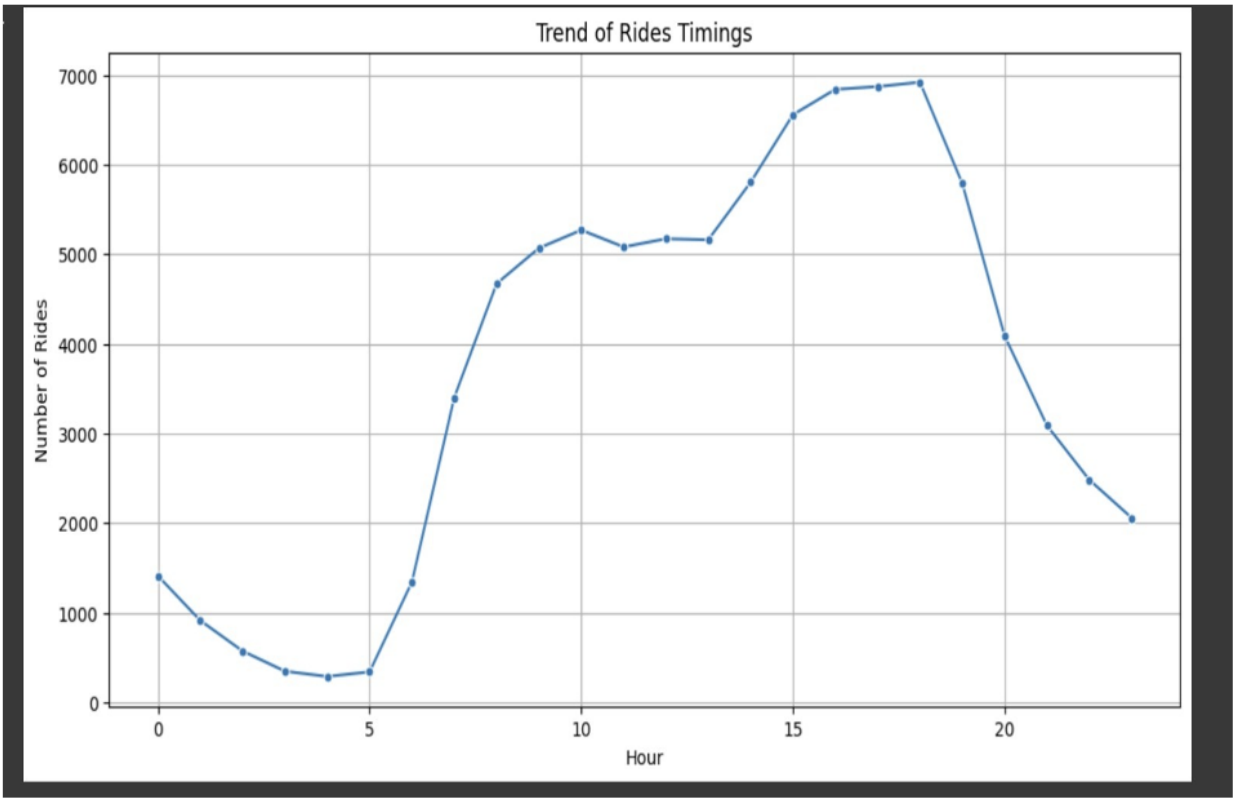
- Load the datasets into Python.
- Converting the parquet data file into csv.
- Loading Green Taxi datasets separately.
- Dataset characteristics, we have taken 3 data sets January-62495 and August-65929 and Taxi Zones-263.
- Initially we have performed the initial data Analysis of to check the datatypes, drop the redundant columns, renaming columns as per data requirements, and merging the data sets , Creating Functions to efficiently convert dates into datetime datatypes, Removing negative outliers by considering data where dropoff_time > pickup_time and Dropping the NA(missing values from dataset). Now our data is ready for analysis.
- We also checked for outliners as we are using the statics analysis the outliners are datapoints which are far away from data sets cluster which causes the difficult in analysis so we remove the outliners using Inter quartile Range concept.
- We have calculated the tip percent based on tip_amount and total_amount and created a new columns of hour and days of week for analysis in finding which hour drivers get most of the rides based on which hour and which day of week.

- We have also created a new dataset by merge the green trip and taxi zone to find the congestion surge charge.

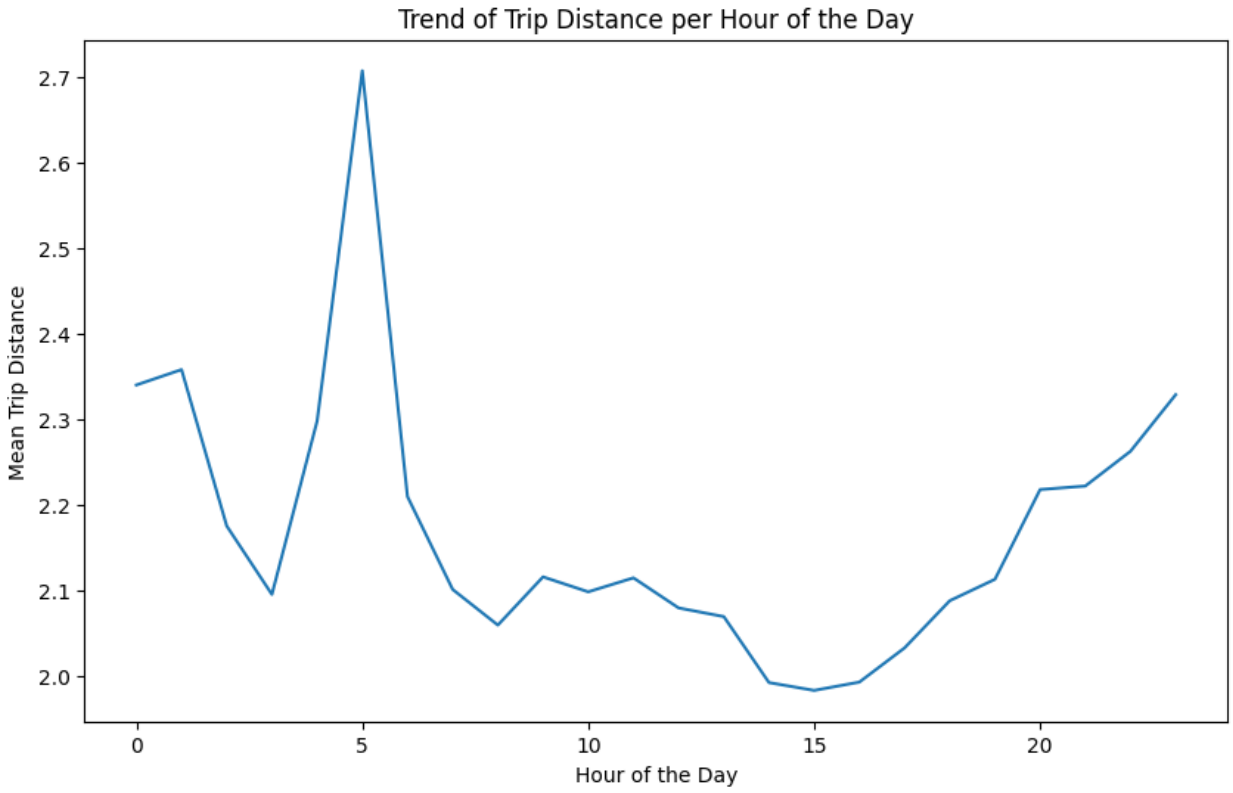
Analysis :



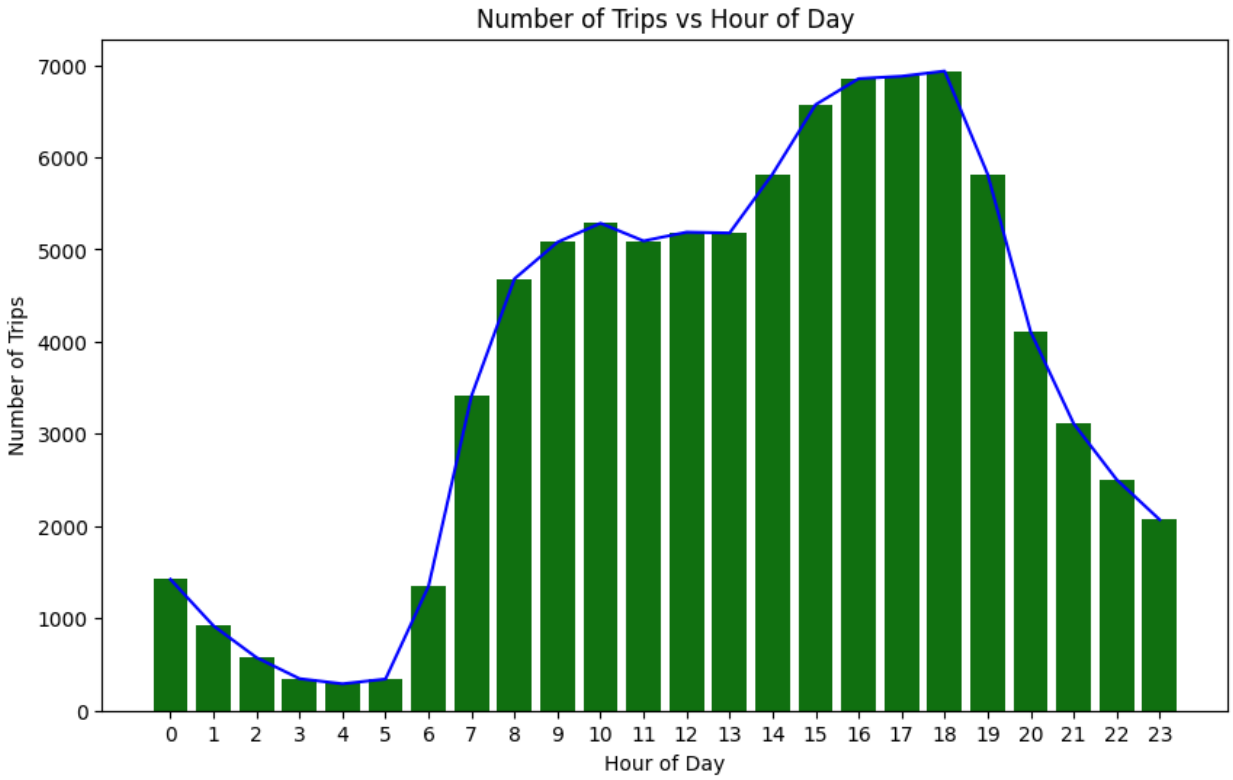
- From the above graph we have found that most of trips were happening within 1-2 miles of range.



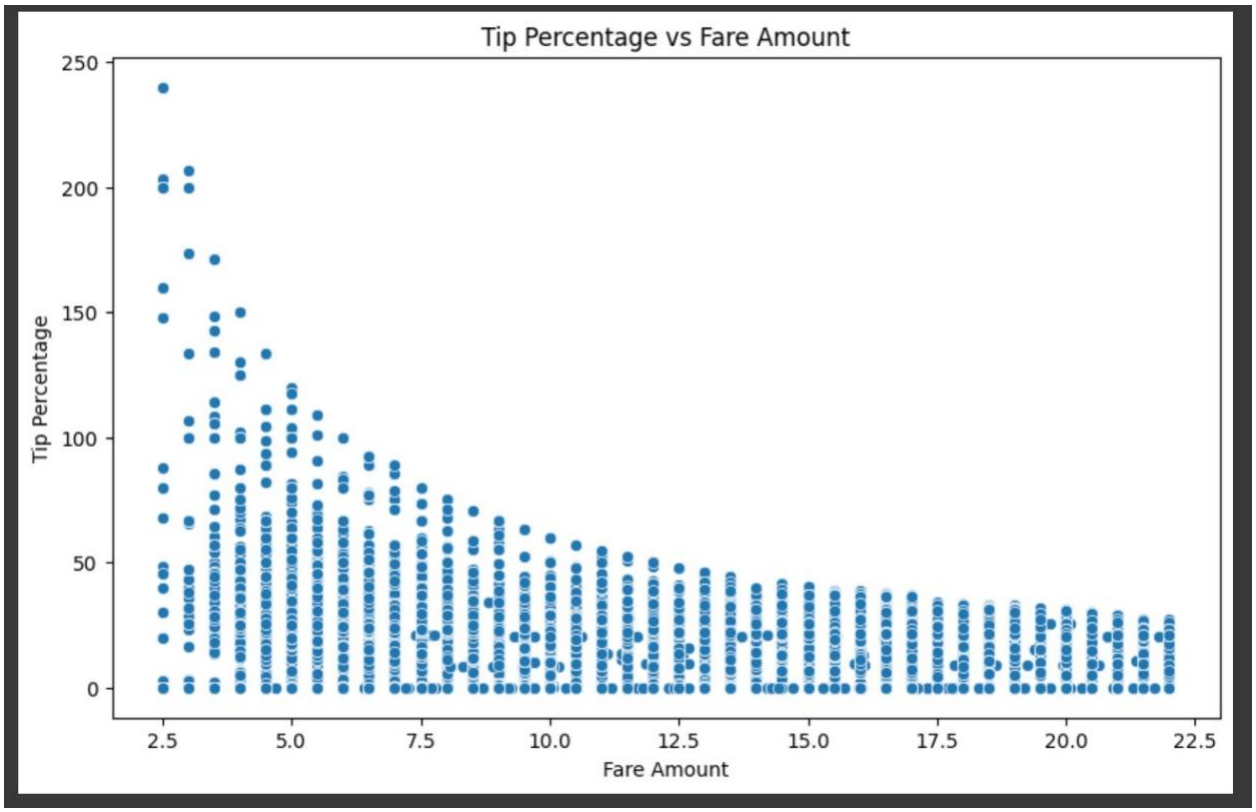
- Here from the graph we are finding the most of rides are happening between the afternoon and evening i.e 15-20.



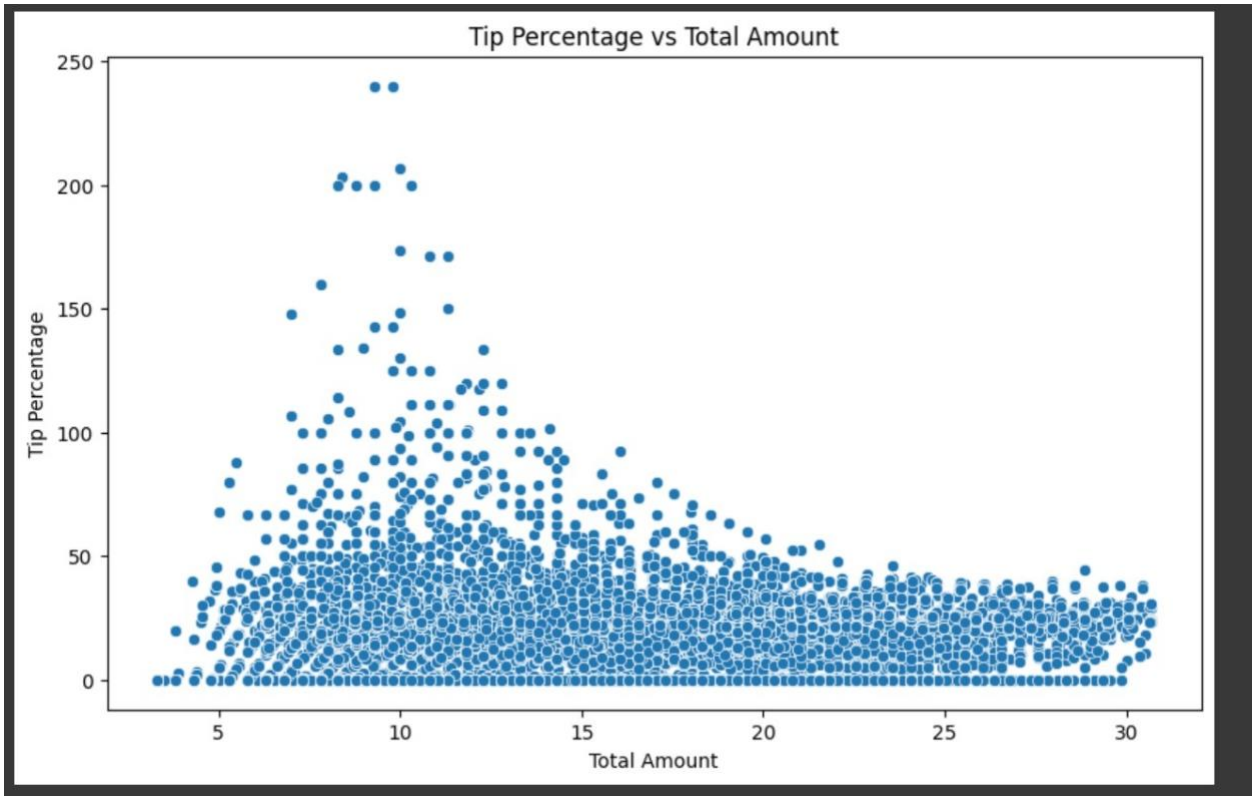
- Here we graph shows that trips gets highest top during the morning hours around 5:00AM



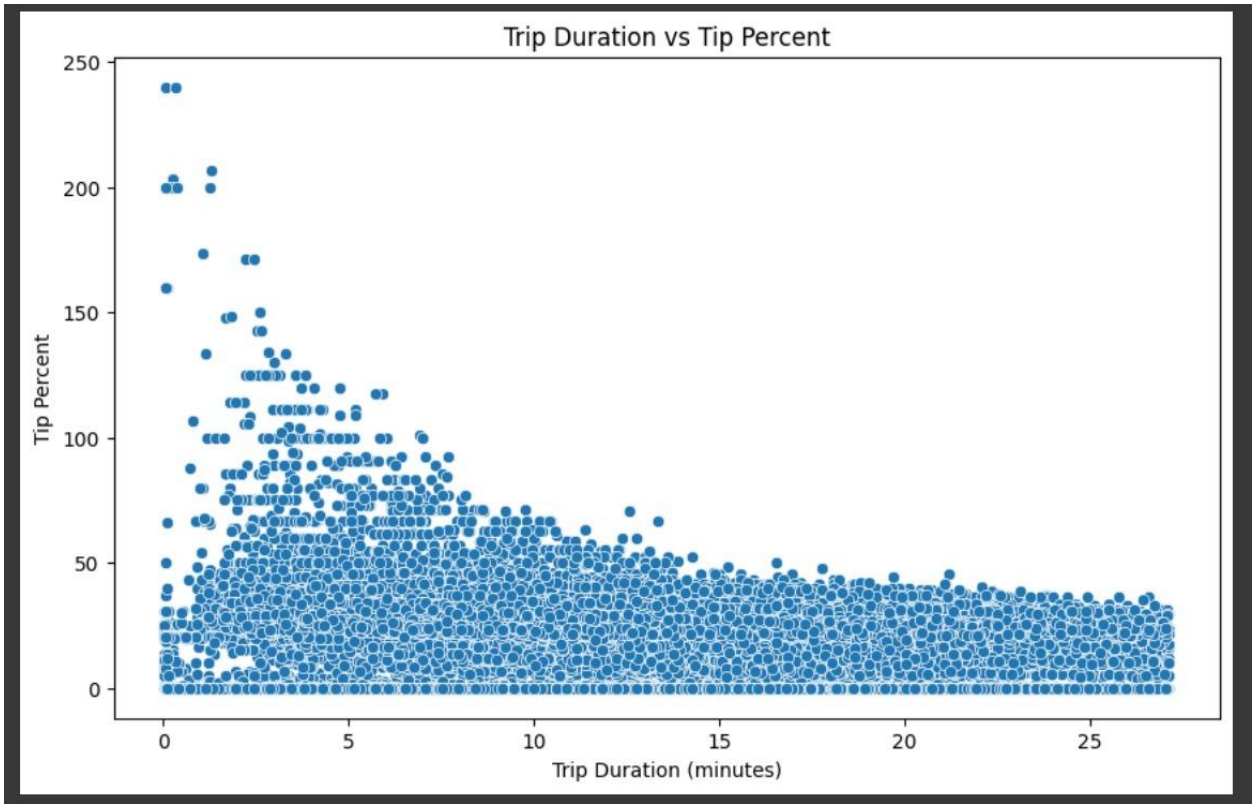
- The above graph gives us that the on which hours drivers get the most of rides booking so we can see that at evening the number trips gets peaks which means of the ride booking are happening at evening than in the morning.



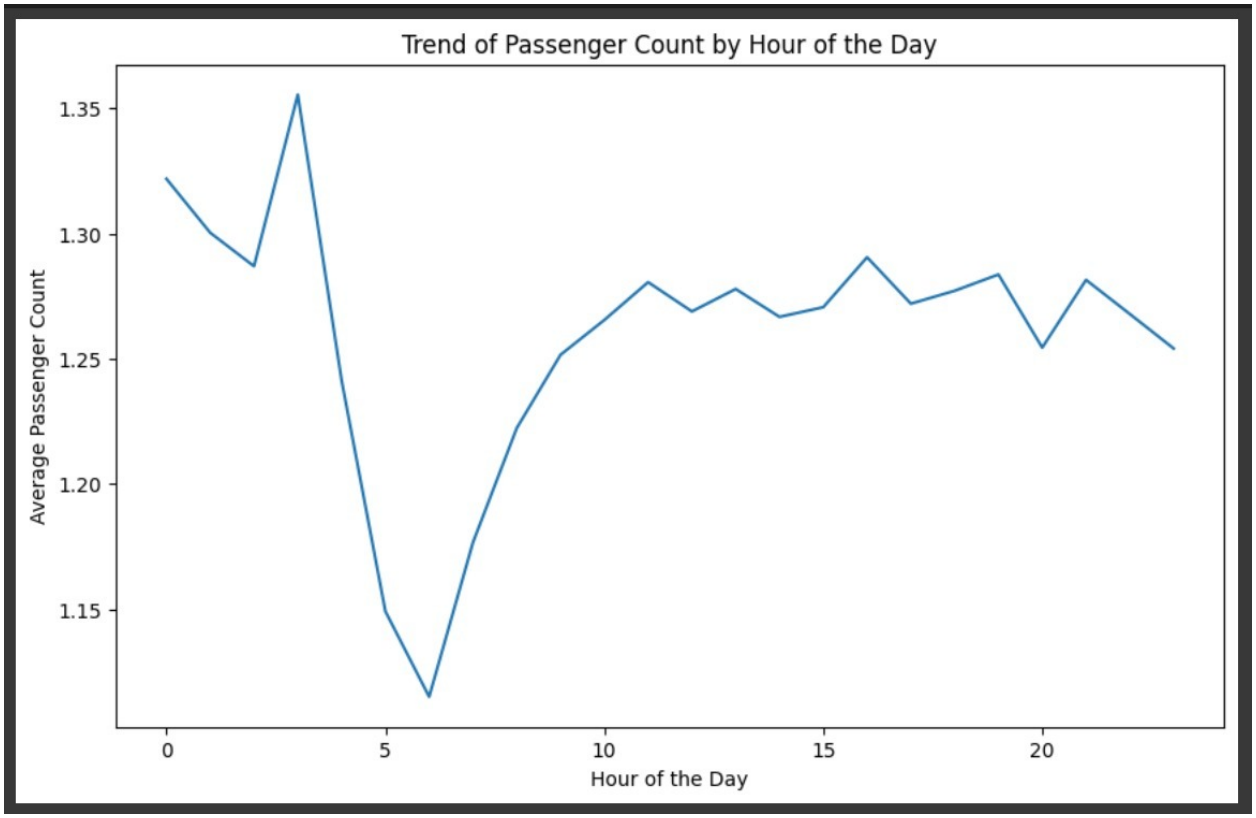
- Here We are find the tip percentage based on the fare amount, The drivers gets the tip mostly if the fare amount is low than the fare amount is higher.



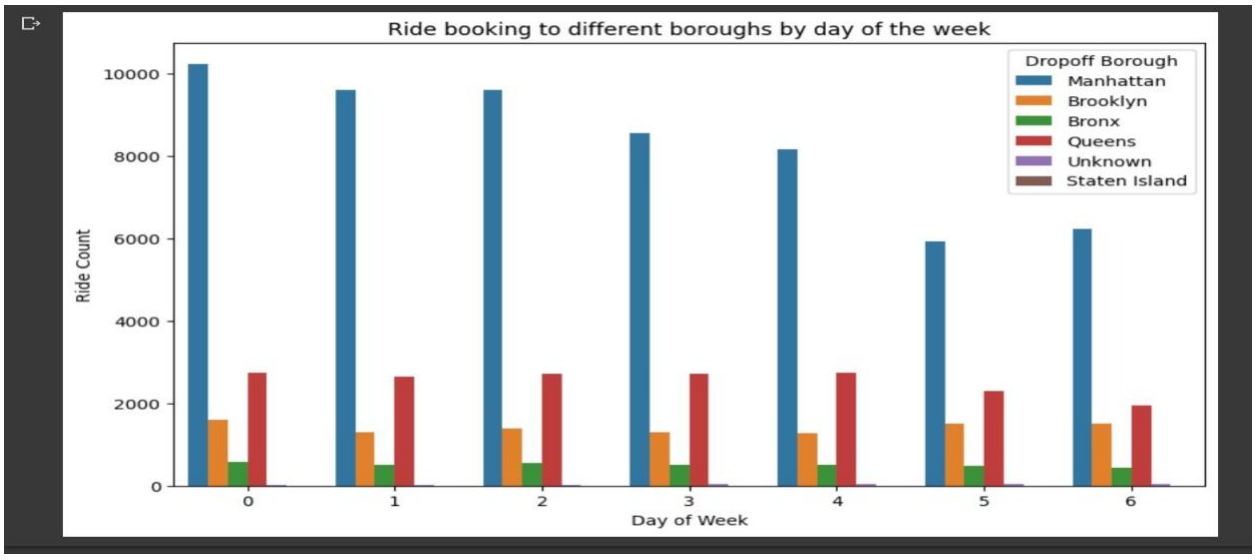
- When the total amount is between 5-15$, the trip percentage has its peaks. The drivers has highest change of getting high tip percentage when the total amount is between 5-15$.



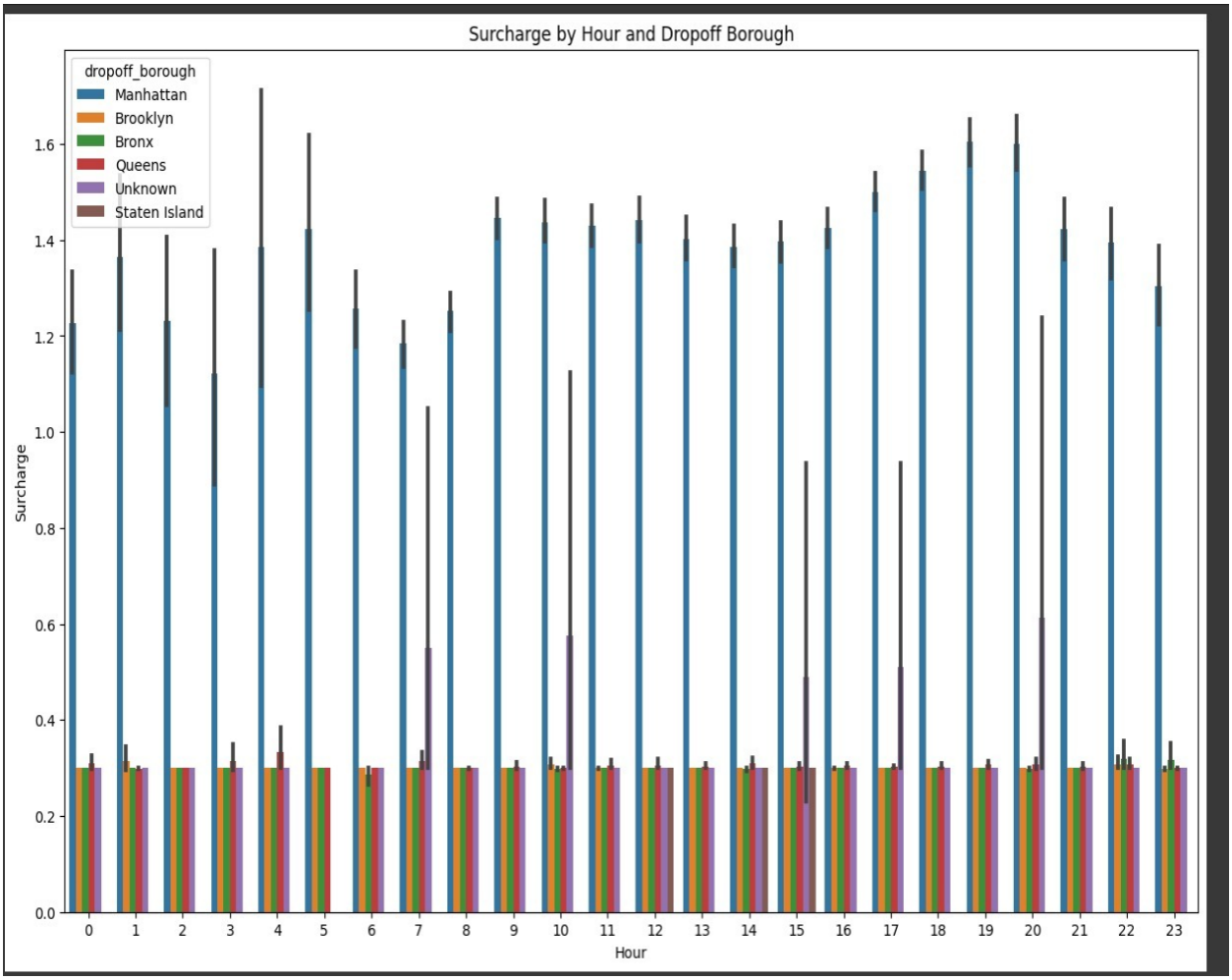
- When the trip duration is between 0-5 minutes. The driver has higher chance of getting the tip percent.



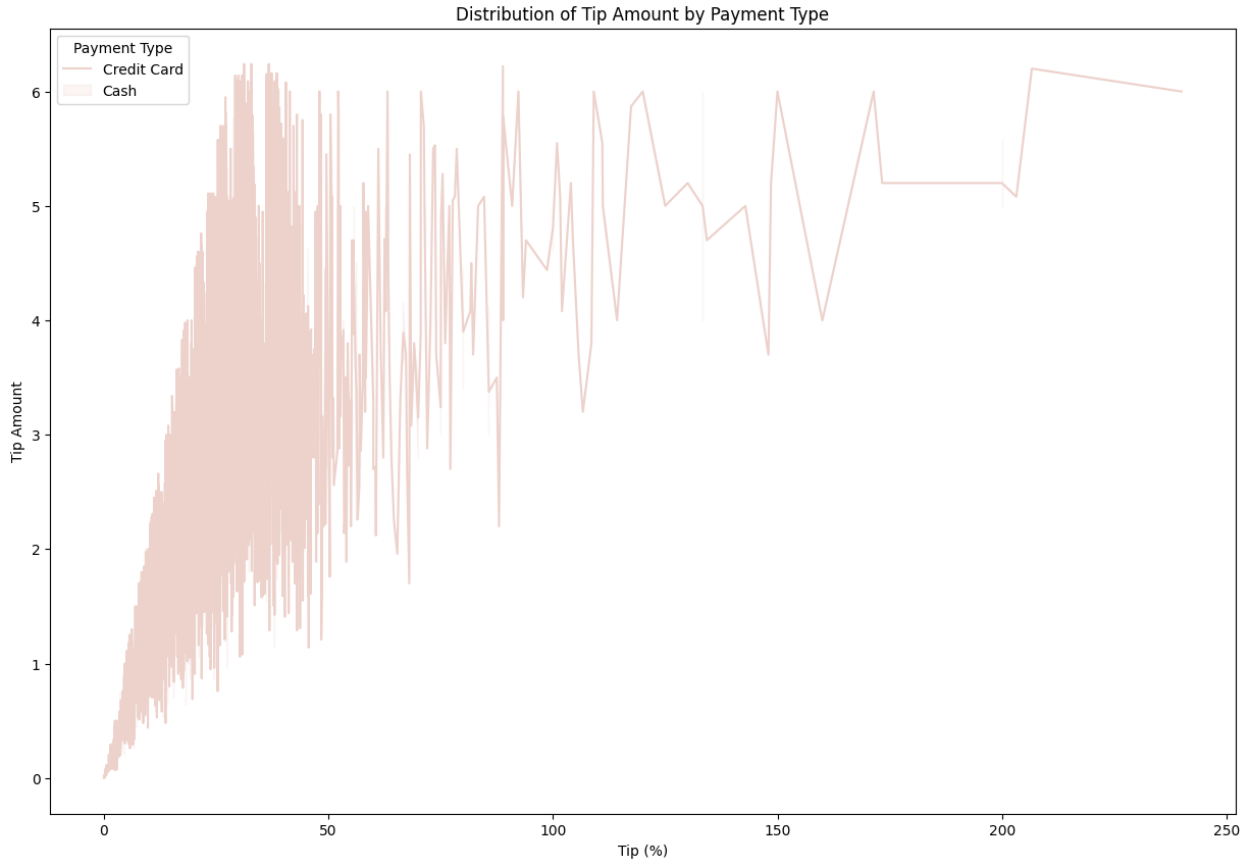
- The average passenger counts is peaks at during early morning.



- The borough that most people travel within or travel to in Green taxis is Manhattan followed by Queens, Brooklyn and Bronx.



- Finding the surcharge by hour based on area For Manhattan, the average surcharge is consistently greater than for any other Borough.
- Early in the morning and again in the evening, the maximum surcharge amount is charged.



- From above graph we have finding what is mode of payments of tip and fare amount. Most of the payment is made through credit cards than cash. Drivers receive most of the payments in credit card mode.

We have used the seaborn function in generating or creating the statical graphs.

Conclusion : From the above exploratory analysis Drivers can benefit from this information by using it, according to the report. This research may be useful for drivers who have to be at their residences early in the morning for pick-up and at corporate locations in New York during business hours. They would then have the chance to take more rides, which would increase their pay. The TLC will be able to develop a number of tactics to make money from green cabs with the aid of this information.

Sravan Kumar Mangalagiri - Worked on Exploratory Analysis and Data Cleaning

Sai Sisira Pathakamuri – Worked on Data cleaning and Exploratory Analysis

Sravan Kumar Mangalagiri

Sai Sisira Pathakamuri